МОЛЕКУЛЯРНАЯ БИОЛОГИЯ, 2025, том 59, № -2, с. 299–308

БИОИНФОРМАТИКА

УДК 577.112.3:577.32

РАЗДЕЛЕНИЕ СТАНДАРТНОГО НАБОРА АМИНОКИСЛОТ НА ГРУППЫ В СООТВЕТСТВИИ С ИХ ЭВОЛЮЦИОННЫМ ВОЗРАСТОМ

© 2025 г. В. М. Ефимов[a, b, c, d, *], К. В. Ефимов[e], В. Ю. Ковалева[b]

[a]*Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск, 630090 Россия*

[b]*Институт систематики и экологии животных Сибирского отделения Российской академии наук, Новосибирск, 630091 Россия*

[c]*Новосибирский государственный университет, Новосибирск, 630090 Россия*

[d]*Томский государственный университет, Томск, 634050 Россия*

[e]*Высшая школа экономики, Москва, 101000 Россия*

**e-mail: vmefimov@gmail.com*

Поступила в редакцию 05.08.2024 г.
После доработки 07.10.2024 г.
Принята к публикации 16.10.2024 г.

Принято считать, что существующий набор протеиногенных аминокислот, кодируемых стандартным генетическим кодом, сформировался поэтапно в ходе эволюции. В большинстве исследований в число ранних входят Ala, Asp, Glu, Gly, Ile, Leu, Pro, Ser, Thr и Val, предположительно имеющие внеземное происхождение. Однако в других работах в качестве консенсусного выбран список ранних аминокислот, в котором место Ile занимает Arg. Мы сравнили различия ранних и поздних аминокислот по совокупности их физико-химических свойств (база данных AAindex). Между бинарными списками с Ile и Arg и каждым АА-индексом вычислены точечно-бисериальный коэффициент корреляции r_{pb}, t- критерий Стьюдента и его достоверности – p-*value*. Поскольку в общей сложности получено 2 × 553 значений p-*value,* проблему множественных сравнений решили с помощью коррекции по Бонферрони и метода Бенджамини–Хохберга. Далее мы использовали метод 2B-PLS, который применяется к двум различным наборам признаков, относящихся к одним и тем же объектам, для поиска общей для них информации. В качестве первого набора взяли бинарные списки по Трифонову (Arg) и Вонгу (Ile), в качестве второго – 553 АА-индекса. Максимальную корреляцию и со списком с Ile, и с Arg (1.0 и 0.8 соответственно) продемонстрировал двоичный АА-индекс CHAM830108, характеризующий способность аминокислоты быть донором заряда: поздние аминокислоты способны быть донорами, а ранние – нет. По-видимому, это обусловлено различиями в условиях, в которых протекала эволюция стандартного набора аминокислот: пребиотических и биотических. Из результатов 2B-PLS-анализа тоже следует, что в списке 10 эволюционно ранних аминокислот Ile выглядит предпочтительнее Arg. Подтверждается выделение последних шести аминокислот (Cys, His, Met, Phe, Trp, Tyr), полученное на основании уменьшения зазора HOMO-LUMO, в отдельный, третий этап эволюции набора стандартных аминокислот. Выявлено компактное расположение на 2B-PLS-плоскости физико-химических свойств трех групп аминокислот, у которых во второй позиции кодонов находятся, соответственно, аденин, тимин и цитозин, а также максимальное рассредоточение аминокислот с гуанином во второй позиции кодонов.

Ключевые слова: ранние и поздние аминокислоты, AAindex, CHAM830108, точечно-бисериальный коэффициент корреляции, поправка Бонферрони, метод Бенджамини–Хохберга, 2B-PLS-анализ

DOI: 10.31857/S0026898425020111, **EDN:** GFVVPE

ВВЕДЕНИЕ

Считается, что набор 20 протеиногенных аминокислот, в настоящее время кодируемых стандартным генетическим кодом, поэтапно расширялся в ходе эволюции, установлен даже примерный порядок их появления в этом наборе [1–10]. В качестве 10 первичных аминокислот указываются Ala, Arg, Asp, Glu, Gly, Leu, Pro, Ser, Thr и Val, которые, возможно, сформировались вне Земли или на ранней Земле до того, как она стала пригодна для жизни (пребиотический синтез). Однако в работах [10–23] в качестве консенсусного приведен несколько отличающийся список ранних аминокислот – Arg

заменен на Ile. В пользу обоих вариантов представлены достаточно серьезные доводы.

Условия, в которых предположительно формировались наборы ранних и поздних аминокислот, кардинально отличаются друг от друга. В первом случае – пребиотический синтез в условиях жесткого космического облучения и солнечного ультрафиолета; во втором – уже созданная первичной жизнью кислородная атмосфера (к которой теперь необходимо приспосабливаться), но с озоновым экраном, защищающим жизнь от излишнего облучения. Поэтому вопрос о конкретном составе этих двух списков является достаточно принципиальным, в первую очередь, для уточнения часто используемого консенсусного порядка появления стандартных аминокислот по Трифонову [2]. Появляется довольно много биологических, эволюционных и даже экологических работ, которые уже опираются на этот порядок, как надежно установленный и общепринятый (эпитет "консенсусный" в немалой степени этому способствует). Поскольку вопрос переходит из теоретической стадии в практическую, требования к надежности этого порядка возрастают.

Набор признаков, характеризующих физико-химические свойства 20 стандартных аминокислот (N = 553), взят из базы данных AAindex [24]. Мы исследовали, насколько различаются по своим физико-химическим свойствам оба списка ранних и поздних аминокислот, надеясь, что это может способствовать прояснению ситуации.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Материалы. Из базы данных AAindex **(https://www.genome.jp/aaindex/)** взяты 553 индекса физико-химических свойств 20 протеиногенных аминокислот (далее AA-индексы) [24]. Дополнительно для 20 аминокислот созданы два новых двоичных признака их принадлежности к эволюционно ранним и поздним. Ранним аминокислотам сопоставлено значение 0, поздним – значение 1. Новым признакам присвоены имена Wong81 (список с Ile = 0) и Trifonov00 (список с Arg = 0) в соответствии с фамилиями их авторов. Оба новых признака добавлены к матрице AA-индексов, фрагмент которой приведен в табл. 1.

Методы. Расчеты велись для каждого признака принадлежности отдельно (табл. 2, 3). Между признаком принадлежности и каждым из 553 AA-индексов вычислены точечно-бисериальный коэффициент корреляции r_{pb} (между количественным признаком и двоичным), *t*-критерий Стьюдента его достоверности по формулам из трехтомника М. Кендалла и А. Стьюарта (разделы 26.34, 26.35) [25], а также сама достоверность *p-value* (табл. 2, 3). В тех же разделах показано, что *t*-критерий Стьюдента достоверности коэффициента корреляции r_{pb} эквивалентен *t* критерию Стьюдента достоверности разности средних двух выборок, на которые двоичный признак фактически разбивает количественную выборку (в нашем случае на 10 ранних и 10 поздних аминокислот для каждого AA-индекса). Заметим, что коэффициенты корреляции между количественным признаком и двоичным, а также между двумя двоичными признаками, как следует из их формул, можно вычислять по обычной формуле линейного коэффициента корреляции Пирсона.

Поскольку в общей сложности получено 2 × 553 значений *p-value,* потребовалось решить проблему множественных сравнений (табл. 2, 3). Для этого все *p-value* упорядочены по возрастанию, каждому присвоен порядковый номер *k,* для каждого вычислены скорректированные значения с поправкой Бонферрони по формуле $p_B = p\text{-}value \times 2 \times 553$ и по методу Бенджамини–Хохберга по формуле $p_{BH} = p_B/k$ [26]. Для метода Бенджамини–Хохберга возможна ситуация, когда p_{BH} для бóльшего *k* меньше p_{BH} для меньшего *k*. В этом случае бóльшее значение заменяется на меньшее (p_{BHadj}).

Современные требования к статистической обработке данных предполагают вычисление, кроме достоверности, еще и размера эффекта, а также научное и практическое обоснование принимаемого решения [27]. Для коэффициента корреляции размером эффекта является его квадрат – доля снимаемой дисперсии признака в линейной регрессии. Мы выбрали в качестве порогового 50%-ный уровень, и из табл. 2 и 3 отобрали только AA-индексы, удовлетворяющие условиям $p_{BHadj} < 0.05$ и $R^2 > 0.5$.

Для описания выявленных индексов мы использовали AAontology [28] – кластеризацию 586 AA-индексов на восемь категорий и 67 подкатегорий, внутри которых индексы статистически и содержательно близки друг к другу и корреляция между которыми не опускается ниже 0.3. В число индексов 586 входит вся база AAindex [24] – 553 индекса. Дополнительные индексы, предложенные авторами AAontology, мы не рассматривали. Категориями являются (в скобках – число подкатегорий): *ASA/Volume* (5), *Composition* (5), *Conformation* (24), *Energy* (9), *Polarity* (6), *Shape* (6), *Structure–Activity* (6) и *Others* (6). Для удобства пользователей авторы предлагают онлайн-сервис: https://www.sciencedirect.com/science/article/pii/S0022836240032 67#b0340.

Таблица 1. Физико-химические свойства 20 протеиногенных аминокислот [24] (фрагмент) и принадлежность к ранним или поздним аминокислотам [1, 11]

AA-индексы	Ala	Arg	Asp	Glu	Gly	Leu	Pro	Ser	Thr	Val	Asn	Cys	Gln	His	Ile	Lys	Met	Phe	Trp	Tyr
Trifonov00	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
Wong81	0	1	0	0	0	0	0	0	0	0	1	1	1	1	0	1	1	1	1	1
ANDN920101	4.35	4.38	4.76	4.29	3.97	4.17	4.44	4.5	4.35	3.95	4.75	4.65	4.37	4.63	3.95	4.36	4.52	4.66	4.7	4.6
ARGP820101	0.61	0.6	0.46	0.47	0.07	1.53	1.95	0.05	0.05	1.32	0.06	1.07	0	0.61	2.22	1.15	1.18	2.02	2.65	1.88
ARGP820102	1.18	0.2	0.05	0.11	0.49	3.23	0.76	0.97	0.84	1.08	0.23	1.89	0.72	0.31	1.45	0.06	2.67	1.96	0.77	0.39
ARGP820103	1.56	0.45	0.14	0.23	0.62	2.93	0.76	0.81	0.91	1.14	0.27	1.23	0.51	0.29	1.67	0.15	2.96	2.03	1.08	0.68
BEGF750101	1	0.52	0.44	0.73	0.35	1	0.06	0.35	0.44	0.82	0.35	0.06	0.44	0.6	0.73	0.6	1	0.6	0.73	0.44
BEGF750102	0.77	0.72	0.65	0.55	0.65	0.83	0.55	0.55	0.83	0.98	0.55	0.65	0.72	0.83	0.98	0.55	0.98	0.98	0.77	0.83
BEGF750103	0.37	0.84	0.97	0.53	0.97	0.53	0.97	0.84	0.75	0.37	0.97	0.84	0.64	0.75	0.37	0.75	0.64	0.53	0.97	0.84
BHAR880101	0.357	0.529	0.511	0.497	0.544	0.365	0.509	0.507	0.444	0.386	0.463	0.346	0.493	0.323	0.462	0.466	0.295	0.314	0.305	0.42
BIGC670101	52.6	109.1	68.4	84.7	36.3	102	73.6	54.9	71.2	85.1	75.7	68.3	89.7	91.9	102	105.1	97.7	113.9	135.4	116.2
BIOV880101	16	−70	−78	−106	−13	145	−20	−70	−38	123	−74	168	−73	50	151	−141	124	189	145	53
BIOV880102	44	−68	−91	−139	−8	108	−36	−60	−54	117	−72	90	−117	47	100	−188	121	148	163	22
BROC820101	7.3	−3.6	−2.9	−7.1	−1.2	20	5.1	−4.1	0.8	3.5	−5.7	−9.2	−0.3	−2.1	6.6	−3.7	5.6	19.2	16.3	5.9
BROC820102	3.9	3.2	−2.8	−7.5	−2.3	15	5.6	−3.5	1.1	2.1	−2.8	−14.3	1.8	2	11	−2.5	4.1	14.7	17.8	3.8
BULH740101	−0.2	−0.12	−0.2	−0.3	0	−2.46	−0.98	−0.39	−0.52	−1.56	0.08	−0.45	0.16	−0.12	−2.26	−0.35	−1.47	−2.33	−2.01	−2.24
BULH740102	0.691	0.728	0.558	0.632	0.592	0.842	0.73	0.594	0.655	0.777	0.596	0.624	0.649	0.646	0.809	0.767	0.709	0.756	0.743	0.743
BUNA790102	4.349	4.396	4.765	4.295	3.972	4.385	4.471	4.498	4.346	4.184	4.755	4.686	4.373	4.63	4.224	4.358	4.513	4.663	4.702	4.604

Таблица 2. Расчет достоверности различий и размера эффекта по АА-индексам (фрагмент) между ранними и поздними аминокислотами по [1] с учетом множественности сравнений (2 × 553 индекса)

Trifonov00	*R*	R^2	t^2	*t*	*p-value*	*k*	p_B	p_{BH}	p_{BHadj}
CHAM830108	0.800	**0.640**	32.0	5.66	2.3E-05	1	*0.0254*	*0.0254*	*0.0141*
NAKH920101	–0.797	**0.636**	31.4	5.60	2.6E-05	2	*0.0283*	*0.0141*	*0.0141*
NAKH920103	–0.783	**0.613**	28.6	5.34	4.4E-05	3	*0.0491*	*0.0164*	*0.0164*
CEDJ970102	–0.764	**0.583**	25.2	5.02	8.9E-05	4	0.0986	*0.0246*	*0.0214*
DAYM780101	–0.758	**0.574**	24.3	4.93	0.00011	5	0.1198	*0.0240*	*0.0214*
JOND920101	–0.754	**0.568**	23.7	4.87	0.00012	6	0.1371	*0.0229*	*0.0214*
NAKH900101	–0.751	**0.564**	23.3	4.83	0.00014	7	0.1499	*0.0214*	*0.0214*
HUTJ700101	0.739	**0.546**	21.6	4.65	0.00020	8	0.2201	*0.0275*	*0.0267*
CEDJ970104	–0.733	**0.538**	21.0	4.58	0.00023	9	0.2577	*0.0286*	*0.0267*
CEDJ970105	–0.732	**0.536**	20.8	4.56	0.00024	10	0.2674	*0.0267*	*0.0267*
NAKH920102	–0.717	**0.514**	19.0	4.36	0.00037	11	0.4141	*0.0376*	*0.0376*
QIAN880129	0.714	**0.509**	18.7	4.32	0.00041	12	0.4526	*0.0377*	*0.0377*
CEDJ970101	–0.699	0.488	17.2	4.14	0.00061	13	0.6752	0.0519	0.0519
NAKH920107	–0.693	0.480	16.6	4.08	0.00071	14	0.7832	0.0559	0.0559
FUKS010110	–0.685	0.470	15.9	3.99	0.00085	15	0.9430	0.0629	0.0629
JUKT750101	–0.679	0.461	15.4	3.92	0.00100	16	1.1099	0.0694	0.0694
PRAM820102	–0.668	0.446	14.5	3.81	0.00129	17	1.4248	0.0838	0.0768
NAKH900102	–0.667	0.445	14.4	3.80	0.00131	18	1.4540	0.0808	0.0768
JUNJ780101	–0.667	0.445	14.4	3.80	0.00132	19	1.4585	0.0768	0.0768
KUMS000102	–0.644	0.415	12.7	3.57	0.00219	20	2.4183	0.1209	0.1209

Примечание. Курсивом выделены достоверные *p-value* (< 0.05) с коррекцией по Бонферрони (*pB*) и методу Бенджамини–Хохберга (*pBH, pBHadj*). Жирным выделены АА-индексы, для которых размер эффекта превышает 50% ($R^2 > 0.5$).

Далее мы использовали метод 2B-PLS, который применяется к двум различным наборам признаков, относящихся к одним и тем же объектам, для поиска общей для обоих наборов информации [29]. На первом шаге ищутся бикомпоненты – две линейные комбинации признаков с максимальной ковариацией между ними, каждая в своем наборе. Потом этот шаг повторяется на подпространствах, оставшихся после “снятия” бикомпонент. Метод аналогичен каноническому корреляционному анализу Хотеллинга, но, в отличие от него, гораздо лучше работает на практике. Имеется во всех современных статпакетах (R, Statistica. PAST и т.д.).

В качестве первого набора мы взяли бинарные списки по Трифонову и Вонгу, в качестве второго – 553 АА-индекса. Нас интересовало, как располагаются аминокислоты относительно друг друга в соответствующем пространстве бикомпонент (рис. 1). Расчеты велись с помощью пакетов Statistica12 [30], PAST4 [31] и Jacobi4 [32].

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

С признаком Trifonov00 достоверную по Бонферрони корреляцию $p_B < 0.05$ демонстрируют три АА-индекса: CHAM830108, NAKH920101 и NAKH920103 (табл. 2). Коррекция по Бонферрони, как известно является самой жесткой [26]. Самым мягким, но тем не менее достаточно адекватным считается метод Бенджамини–Хохберга, и в табл. 2 по нему с достоверностью $p_{BHadj} < 0.05$ набирается 12 АА-индексов из 553. Ни один из них не вышел на уровень $p_{BHadj} < 0.01$. Индексов с размером эффекта $R^2 > 0.5$ оказалось тоже 12, т.е., оба набора совпали (табл. 2).

В соответствии с AAontology, АА-индекс CHAM830108 [30] попадает в категорию *Energy*, подкатегорию *Charge*. Для краткости будем писать *Energy* (*Charge*). Согласно авторскому описанию АА-индекса CHAM830108 [33], он характеризует способность аминокислоты быть донором заряда. При сопоставлении этой способности с перечнем стандартных амино-

Таблица 3. Расчет достоверности различий и размера эффекта по АА-индексам (фрагмент) между ранними и поздними аминокислотами по [11] с учетом множественности сравнений (553 индекса)

Wong81	R	R^2	t^2	t	p	k	p_B	p_{BH}	p_{BHadj}
CHAM830108	1	**1**	–	–	0	1	*0*	*0*	*0*
NAKH920103	–0.784	**0.615**	28.8	5.36	4.30E-05	2	*0.0472*	*0.0236*	*0.0163*
KARS160115	0.780	**0.609**	28.0	5.29	5.00E-05	3	0.0550	*0.0183*	*0.0163*
CEDJ970102	–0.769	**0.591**	26.0	5.10	7.50E-05	4	0.0833	*0.0208*	*0.0163*
KARS160114	0.767	**0.588**	25.6	5.06	8.10E-05	5	0.0895	*0.0179*	*0.0163*
JOND920101	–0.764	**0.583**	25.2	5.02	8.90E-05	6	0.0984	*0.0164*	*0.0163*
KARS160116	0.758	**0.575**	24.4	4.94	0.00011	7	0.1178	*0.0168*	*0.0163*
FUKS010110	–0.755	**0.570**	23.9	4.89	0.00012	8	0.1308	*0.0163*	*0.0163*
CEDJ970104	–0.748	**0.560**	22.9	4.78	0.00015	9	0.1648	*0.0183*	*0.0183*
KUMS000102	–0.743	**0.552**	22.2	4.71	0.00017	10	0.1935	*0.0193*	*0.0193*
DAYM780101	–0.739	**0.546**	21.6	4.65	0.00020	11	0.2207	*0.0201*	*0.0201*
KUMS000101	–0.733	**0.537**	20.9	4.57	0.00024	12	0.2604	*0.0217*	*0.0204*
NAKH900101	–0.732	**0.537**	20.8	4.56	0.00024	13	0.2656	*0.0204*	*0.0204*
CEDJ970101	–0.718	**0.516**	19.2	4.38	0.00036	14	0.3999	*0.0286*	*0.0254*
JUKT750101	–0.718	**0.515**	19.1	4.37	0.00037	15	0.4066	*0.0271*	*0.0254*
HUTJ700102	0.714	**0.510**	18.8	4.33	0.00040	16	0.4449	*0.0278*	*0.0254*
PRAM820102	–0.713	**0.508**	18.6	4.31	0.00042	17	0.4640	*0.0273*	*0.0254*
MCMT640101	0.712	**0.507**	18.5	4.30	0.00043	18	0.4749	*0.0264*	*0.0254*
CEDJ970103	–0.710	**0.504**	18.3	4.28	0.00045	19	0.4989	*0.0263*	*0.0254*
CHAM830106	0.709	**0.503**	18.2	4.27	0.00046	20	0.5085	*0.0254*	*0.0254*

Примечание. Курсивом выделены достоверные *p-value* (< 0.05) с коррекцией по Бонферрони (*pB*) и методу Бенджамини–Хохберга (*pBH, pBHadj*). Жирным выделены АА-индексы, для которых размер эффекта превышает 50% ($R^2 > 0.5$).

кислот получается, что 9 из 10 поздних аминокислот по списку Trifonov00 (за исключением Ile) обладают этой способностью, а 9 из 10 ранних (за исключением Arg) – нет. Из 11 остальных индексов 9 попадают в *Composition* (*AA Composition*) – частоты аминокислот в белках различных локаций, например, в мембранных, межклеточных и т.д. [34–37]. Судя по отрицательным корреляциям всех композиционных признаков с Trifonov00, суммарные частоты поздних аминокислот в белках практически всех локаций уступают суммарным частотам ранних.

Остаются HUTJ700101 [38] и QIAN880129 [39]. HUTJ700101 относится к *Energy*, QIAN880129 – к *Conformation*. Но оба имеют подкатегории (*Unclassified*), т.е., не имеют содержательной интерпретации. Надо отметить, что в AAontology из 586 АА-индексов более 40 остались не классифицированными на подкатегории.

Строго говоря, индексы категории *Composition* вообще не являются физико-химическими свойствами, так как характеризуют долю аминокислоты среди других аминокислот, а не саму аминокислоту. Физико-химические свойства все же должны быть измеримыми свойствами самих аминокислот. Конечно, разница в частотах тоже имеет какую-то вескую причину и должна анализироваться статистически, но самоинтерпретируемой не является.

С признаком Wong81 достоверную по Бонферрони корреляцию $p_B < 0.05$ демонстрируют всего два уже известных нам АА-индекса: CHAM830108 и NAKH920103 (табл. 3). Однако по методу Бенджамини–Хохберга с достоверностью $p_{BHadj} < 0.05$ и размером эффекта более 50% ($R^2 > 0.5$) набирается уже 20 АА-индексов (табл. 3).

Корреляция между Wong81 и CHAM830108 равна 1.0 (табл. 3). Следовательно, все 10 поздних аминокислот по списку Wong81 обладают способностью быть донором заряда, а все 10 ранних – нет.

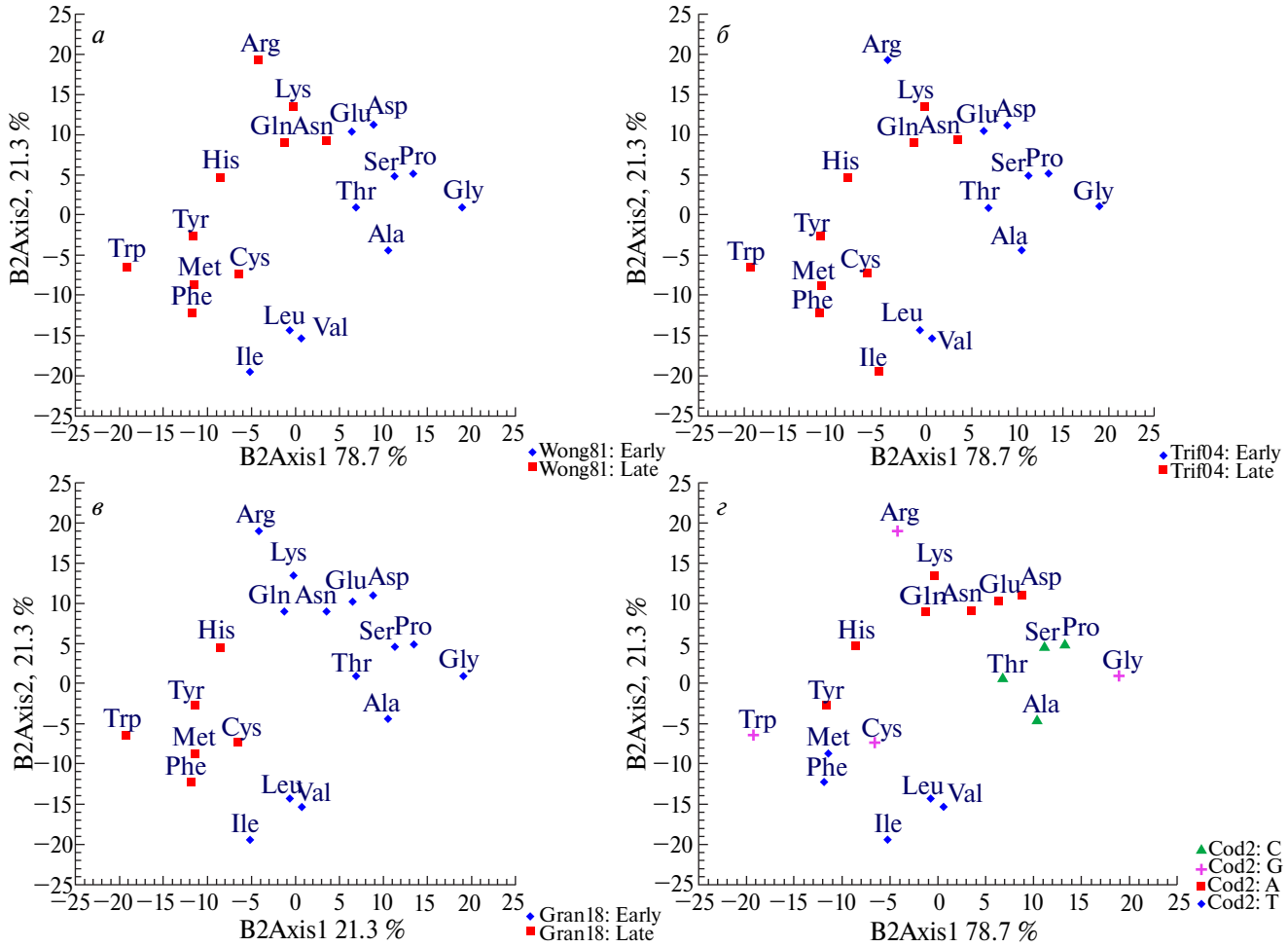

Рис. 1. Конфигурация аминокислот на плоскости двух первых бикомпонент блока физико-химических свойств. Показаны разбиения на ранние и поздние аминокислоты: по Вонгу [11] (*а*); по Трифонову [2] (*б*); по зазору HOMO-LUMO [6] (*в*) и классификация аминокислот по второй позиции кодонов (*г*).

АА-индексов, не классифицированных на подкатегории, нет. Семь АА-индексов: CHAM830108 *Energy* (*Charge*) и NAKH920103, CEDJ970102, JOND920101, CEDJ970104, DAYM780101, NAKH900101 *Composition* (*AA Composition*) входят в предыдущий набор. Из оставшихся 13 в *Composition* (*AA Composition*) также входят FUKS010110 [40], KUMS000102, KUMS000101 [41], CEDJ970101 [35], JUKT750101 [42], и в *Composition* (*Membrane proteins* (*MPs*)) – CEDJ970103 [35].

По сравнению с предыдущим набором появляются сразу пять АА-индексов новой категории: *Shape* (*Side chain length*) – KARS160115, KARS160114, KARS160116[1] (минимальный, средневзвешенный и максимальный эксцентриситеты, основанные на атомном номере) [43], CHAM830106 (количество связей в самой длинной цепочке) [33] и *Shape* (*Shape and Surface*) – PRAM820102 (наклон в регрессионном анализе площади контакта растворителя и пространственного положения) [44]. Два оставшихся индекса: HUTJ700102 *Energy* (*Entropy*) (абсолютные энтропии аминокислот) [38] и MCMT640101 *Polarity* (*Amphiphilicity*) (преломляющая способность) [45] относятся к другим категориям.

Возможным объяснением достоверных корреляций первых шести индексов со списком Wong81, особенно CHAM830106, является то, что все они характеризуют бóльшую или меньшую разветвленность аминокислот. Разветвленность аминокислот необходима для эффективного фолдинга белка, а “способность к фолдингу играла важную роль в определении того, какие (ранние) аминокислоты в конечном итоге стали частью канонического алфавита”

[1] Примечание. В базе AAindex обнаружена опечатка в цитировании источника, отразившаяся в названиях индексов. Вместо KARS должно быть KAKS.

[23]. Таким образом, подтверждается версия о бóльшем соответствии списка Wong81 разбиению аминокислот на ранние и поздние.

Вернемся к Wong81. АА-индекс CHAM830108 и Wong81 полностью совпадают (табл. 3). У двоичных признаков, естественно, несколько больше шансов совпасть, чем у количественных. Однако даже если все количественные признаки превратить в двоичные через разбиение каждого признака на 10 меньших и 10 бóльших значений, то ни один из остальных 552 такого совпадения не показывает. Только АА-индекс CHAM830108 полностью эквивалентен разбиению набора аминокислот на ранние и поздние по Вонгу [11]. На рис. 1*а* ранние аминокислоты по списку Wong81 расположены достаточно компактно и более логично, в отличие от списка Trifonov00 на рис. 1*б*. Более того, список Wong81 полностью совпадает с “миллеровским списком”, который часто приводится и в работах Э.Н. Трифонова как предположительно ранний, причем такой, который обязательно надо принимать во внимание при построении общего списка. Несмотря на это, Э.Н. Трифонов оказывает явное предпочтение аргинину, поскольку аргинин имеет шесть кодонов, а изолейцин – только три. По мнению Э.Н. Трифонова только ранние аминокислоты могут иметь шесть кодонов.

Конфигурация аминокислот на плоскости первых двух бикомпонент блока 553 физико-химических свойств, полученная с помощью 2B-PLS-анализа, совсем не выглядит случайной (рис. 1*а*–*г*). Близкие по физико-химическим свойствам аминокислоты явно тяготеют друг к другу.

Если обратить внимание на вторую позицию в кодонах аминокислот (рис. 1*г*), то группы аденина (Asp, Glu, Asn, Gln, Lys, His, Tyr), тимина (Leu, Val, Ile, Met, Phe) и цитозина (Ala, Ser, Pro, Thr) расположены достаточно компактно, в отличие от группы гуанина (Gly, Arg, Cys, Trp), в которой три из четырех аминокислот заняли максимально далекие друг от друга положения. Группы аминокислот с аденином и тимином во второй позиции кодонов практически противоположны по знаку второй бикомпоненты. Заметно также отсутствие среди поздних тех аминокислот, которые во второй позиции генетического кода содержат цитозин.

В [6] обращено внимание на то, что среди факторов, потенциально значимых для эволюционного выбора аминокислот, не рассматривали квантово-химические свойства, хотя они и составляют основу химической реакционной способности: “Анализируя орбитальные энергии канонических протеиногенных АА и селеноцистеина, мы обнаружили, что энергетическое расстояние между самой высокой занятой молекулярной орбиталью HOMO и самой низкой незанятой LUMO, зазор HOMO-LUMO, проявляет заметную закономерность, связанную с временны́м появлением АА в стандартном генетическом коде. Поздние дополнения к генетическому коду ([5, 18])[2] имеют существенно меньшие зазоры, чем все ранние дополнения, такие как закодированные пребиотические АА, обнаруженные в метеорите Мерчисон ([46]) или извлеченные из эксперимента Юри–Миллера ([47]). Исходя из пограничной теории молекулярных орбиталей ([48]), зазор HOMO-LUMO является одним из наиболее широко применяемых теоретических предикторов химической реакционной способности, в частности, отражая кинетическую стабильность соединения по отношению к реакциям, включающим перенос или перегруппировку электронов ([49, 50])”.

Далее авторы приходят к выводу, что шесть самых поздних аминокислот (Cys, His, Met, Phe, Trp, Tyr) отличаются от 14 более ранних появлением на третьем этапе эволюции по Трифонову – захвате кодонов. По сути, предлагается считать ранними не 10, а 14 аминокислот. Нетрудно заметить, что именно эти шесть самых поздних аминокислот занимают достаточно обособленное положение в левой части рис. 1 (рис. 1*в*). В то же время, если судить по рис. 1 в цитируемой статье, то по зазору HOMO-LUMO аргинин и лизин являются самыми поздними из 14 ранних, т.е., не попадают в 10 самых ранних, а изолейцин, без сомнения, попадает. При этом, в [6] отмечено, что из списка свойств, представленных в базе AAindex, ни одно не относится к таким, как энергия орбиталей или возбужденное состояние.

По нашему мнению, авторы [6] не совсем правы. Как минимум, один признак, имеющий некоторое отношение к энергии орбиталей или возбужденным состояниям, в базе AAindex все же есть, и это все тот же CHAM830108 – способность аминокислоты быть донором заряда. Изначально этот признак был даже количественным (IP, ионизационный потенциал), но потом супруги Чартон установили порог – 10 эВ – и предложили использовать его как индикаторный [33]. В этом качестве он и вошел во все справочники. (Надо сказать, что в 80-е годы достаточно много количественных признаков химических соединений перевели в индикаторные, и авторы [51] отмечают, что вследствие этого и регрессионные уравнения в QSAR упростились и яснее стал их физико-химический смысл.) Судя по совпадению результатов, полученных авторами [6], отно-

[2] Номера источников в оригинале заменены на номера этих же источников в списке литературы к данной статье.

сительно эволюционного порядка аминокислот на основании зазора HOMO-LUMO, с аналогичными результатами, полученными нами при использовании двоичного признака CHAM830108, характеризующего способность аминокислоты быть донором заряда (рис. 1*г*), у них действительно есть много общего.

ЗАКЛЮЧЕНИЕ

В результате статистического анализа различий физико-химических свойств поздних и ранних аминокислот обнаружено, что только АА-индекс CHAM830108 [33] полностью совпадает со списком Вонга [11]. АА-индекс CHAM830108 характеризует способность аминокислоты быть донором заряда: поздние аминокислоты способны быть донорами, а ранние – нет. Вероятно, это обусловлено различиями в условиях, в которых протекали возникновение и эволюция аминокислот, пребиотических и биотических. Из результатов 2B-PLS-анализа следует, что в списке 10 эволюционно ранних аминокислот изолейцин выглядит предпочтительнее аргинина. Подтверждается выделение последних шести аминокислот (Cys, His, Met, Phe, Trp, Tyr), полученное на основании уменьшения зазора HOMO-LUMO [6], в отдельный, третий этап эволюции набора стандартных аминокислот. Выявлено компактное расположение на 2B-PLS-плоскости физико-химических свойств трех групп аминокислот, у которых во второй позиции кодонов находятся, соответственно, аденин, тимин и цитозин, а также максимальное рассредоточение аминокислот с гуанином во второй позиции кодонов.

Работа выполнена при поддержке Программы фундаментальных научных исследований государственных академий наук FWNR-2022-0019 Института цитологии и генетики СО РАН и FWGS-2021-0002 Института систематики и экологии животных СО РАН.

Данная работа выполнена без привлечения человека или животных в качестве объекта исследований.

Авторы заявляют об отсутствии конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

1. Trifonov E.N. (2000) Consensus temporal order of amino acids and evolution of the triplet code. *Gene*. **261**(1), 139–151.
2. Trifonov E.N. (2004) The triplet code from first principles. *J. Biomol. Struct. Dynamics*. **22**(1), 1–11.
3. Sobolevsky Y., Trifonov E.N. (2005) Conserved sequences of prokaryotic proteomes and their compositional age. *J. Mol. Evol*. **61**, 591–596.
4. Jordan I.K., Kondrashov F.A., Adzhubei I.A., Wolf Y.I., Koonin E.V., Kondrashov A.S., Sunyaev S. (2005) A universal trend of amino acid gain and loss in protein evolution. *Nature*. **433**(7026), 633–638.
5. Trifonov E.N. (2009) The origin of the genetic code and of the earliest oligopeptides. *Res. Microbiol*. **160(**7), 481–486.
6. Granold M., Hajieva P., Toşa M.I., Irimie F.D., Moosmann B. (2018) Modern diversification of the amino acid repertoire driven by oxygen. *Proc. Natl. Acad. Sci. USA*. **115**(1), 41–46.
7. Demongeot J., Seligmann H. (2019) Spontaneous evolution of circular codes in theoretical minimal RNA rings. *Gene*. **705**, 95–102.
8. Seligmann H. (2020) First arrived, first served: competition between codons for codon–amino acid stereochemical interactions determined early genetic code assignments. *Sci. Nature*. **107**(3), 20.
9. Saralov A.I. (2021) Factors in protobiomonomer selection for the origin of the standard genetic code. *Acta Biotheoretica*. **69**(4), 745–767.
10. Zhao M., Ding R., Liu Y., Ji Z., Zhao Y. (2022) Determination of the amino acid recruitment order in early life by genome-wide analysis of amino acid usage bias. *Biomolecules*. **12**(2), 171.
11. Wong J.T.F. (1981) Coevolution of genetic code and amino acid biosynthesis. *Trends Biochem. Sci*. **6**, 33–36.
12. Brooks D.J., Fresco J.R., Singh M. (2004) A novel method for estimating ancestral amino acid composition and its application to proteins of the Last Universal Ancestor. *Bioinformatics*. **20**(14), 2251–2257.
13. Wong J.T.F. (2005) Coevolution theory of the genetic code at age thirty. *BioEssays*. **27**(4), 416–425.
14. Doi N., Kakukawa K., Oishi Y., Yanagawa H. (2005) High solubility of random sequence proteins consisting of five kinds of primitive amino acids. *Protein Eng. Des. Sel*. **18**(6), 279–284.
15. Trifonov E.N. (2008) Tracing life back to elements. *Physics Life Rev*. **5**(2), 121–132.
16. Higgs P.G., Pudritz R.E. (2009) A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. *Astrobiology*. **9**(5), 483–490.
17. McDonald G.D., Storrie–Lombardi M.C. (2010) Biochemical constraints in a protobiotic earth devoid of basic amino acids: the “BAA (–) world”. *Astrobiology*. **10**(10), 989–1000.
18. Longo L.M., Blaber M. (2012) Protein design at the interface of the pre-biotic and biotic worlds. *Arch. Biochem. Biophys*. **526**(1), 16–21.
19. Longo L.M., Lee J., Blaber M. (2013) Simplified protein design biased for prebiotic amino acids yields a foldable, halophilic protein. *Proc. Natl. Acad. Sci. USA*. **110**(6), 2135–2139.
20. Doig A.J. (2017) Frozen, but no accident why the 20 standard amino acids were selected. *FEBS J*. **284**(9), 1296–1305.

21. Koonin E.V., Novozhilov A.S. (2017) Origin and evolution of the universal genetic code. *Annu. Rev. Genet.* **51**(1), 45–62.
22. Fried S.D., Fujishima K., Makarov M., Cherepashuk I., Hlouchova K. (2022) Peptides before and during the nucleotide world: an origins story emphasizing cooperation between proteins and nucleic acids. *J. R. Soc. Interface*. **19**(187), 20210641.
23. Makarov M., Sanchez Rocha A.C., Krystufek R., Cherepashuk I., Dzmitruk V., Charnavets T., Hlouchova K. (2023) Early selection of the amino acid alphabet was adaptively shaped by biophysical constraints of foldability. *J. Am. Chem. Soc*. **145**(9), 5320–5329.
24. Kawashima S., Pokarowski P., Pokarowska M., Kolinski A., Katayama T., Kanehisa M. (2007) AAindex: amino acid index database, progress report 2008. *Nucl. Acids Res*. **36**(suppl_1), D202–D205.
25. Кендалл М., Стьюарт А. (1973) *Статистические выводы и связи*. М.: Наука, 900 с.
26. Наркевич А.Н., Виноградов К.А., Гржибовский А.М. (2020) Множественные сравнения в биомедицинских исследованиях: проблема и способы решения. *Экология человека*. **10**, 55–64.
27. Wasserstein R.L., Schirm A.L., Lazar N.A. (2019) Moving to a world beyond "$p< 0.05$". *Am. Statistician*. **73**(suppl. 1), 1–19.
28. Breimann S., Kamp F., Steiner H., Frishman D. (2024) AAontology: an ontology of amino acid scales for interpretable machine learning. *J. Mol. Biol*. 168717.
29. Rohlf F.J., Corti M. (2000) The use of partial least-squares to study covariation in shape. *Systematic Biol.* **49**, 740–753.
30. Hill T., Lewicki P. (2006) *Statistics: methods and applications: a comprehensive reference for science, industry, and data mining*. Tulsa, Okla., UK: StatSoft Ltd. 719 p. ISBN: 9781884233593
31. Hammer Ø., Harper D.A.T., Ryan P.D. (2001) PAST: paleontological statistics software package for education and data analysis. *Palaeontologia Electronica*. **4**, 1–9.
32. Polunin D., Shtaiger I., Efimov V. (2019) JACOBI4 software for multivariate analysis of biological data. *bioRxiv*. 803684.
33. Charton M., Charton B.I. (1983) The dependence of the Chou–Fasman parameters on amino acid side chain structure. *J. Theor. Biol*. **102**(1), 121–134.
34. Nakashima H., Nishikawa K. (1992) The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins. *FEBS Lett*. **303**(2–3), 141–146.
35. Cedano J., Aloy P., Perez–Pons J.A., Querol E. (1997) Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol*. **266**(3), 594–600.
36. Dayhoff M., Schwartz R., Orcutt B. (1978) 22 a model of evolutionary change in proteins. *Atlas Protein Sequence Struct*. **5**, 345–352.
37. Jones D.T., Taylor W.R., Thornton J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*. **8**(3), 275–282.
38. Hutchens J.O. (1970) Heat capacities, absolute entropies, and entropies of formation of amino acids and related compounds. In: *Handbook of Biochemistry*. 2nd ed. Ed. Sober H.A. Cleveland, Ohio: Chem. Rubber Co., B60–B61.
39. Qian N., Sejnowski T.J. (1988) Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol*. **202**(4), 865–884.
40. Fukuchi S., Nishikawa K. (2001) Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. *J. Mol. Biol*. **309**(4), 835–843.
41. Kumar S., Tsai C.J., Nussinov R. (2000) Factors enhancing protein thermostability. *Protein Engineering*. **13**(3), 179–191.
42. Jukes T.H., Holmquist R., Moise H. (1975) Amino acid composition of proteins: selection against the genetic code. *Science*. **189**(4196), 50–51.
43. Kakraba S., Knisley D. (2016) A graph theoretic model of single point mutations in the cystic fibrosis transmembrane conductance regulator. *J. Adv. Biotech*. **6**, 780–786.
44. Prabhakaran M., Ponnuswamy P.K. (1982) Shape and surface features of globular proteins. *Macromolecules*. **15**(2), 314–320.
45. McMeekin T.L., Groves M.L., Hipp N.J. (1964) Refractive indices of amino acids, proteins, and related substances. In: *Amino Acids and Serum Proteins*. Ed. Stekol J.A. Washington: Advances in Chemistry, Am. Chem. Soc., **44**, pp. 54–66.
46. Cronin J.R., Pizzarello S. (1983) Amino acids in meteorites. *Adv. Space Res*. **3**, 5–18.
47. Miller S.L. (1953) A production of amino acids under possible primitive earth conditions. *Science*. **117**, 528–529.
48. Fukui K. (1982) Role of frontier orbitals in chemical reactions. *Science*. **218**, 747–754.
49. Pearson R.G. (1986) Absolute electronegativity and hardness correlated with molecular orbital theory. *Proc. Nat. Acad. Sci. USA*. **83**, 8440–8441.
50. Aihara J. (1999) Reduced HOMO-LUMO gap as an index of kinetic stability for polycyclic aromatic hydrocarbons. *J. Phys. Chem. A*. **103**, 7487–7495.
51. Selassie C.D., Verma R.P. (2003) History of quantitative structure-activity relationships. *Burger's Med. Chem. Drug Discov*. **1**, 1–48.

Dividing of the Standard Set of Amino Acids into Groups According to Their Evolutionary Age

© 2025 V. M. Efimov[1, 2, 3, 4, *], K. V. Efimov[5], V. Yu. Kovaleva[2]

[1]*Institute of Cytology and Genetics, Siberian Branch, Russian Academy of Sciences, Novosibirsk, 630090 Russia*
[2]*Institute of Animal Systematics and Ecology, Siberian Branch, Russian Academy of Sciences, Novosibirsk, 630091 Russia*
[3]*Novosibirsk State University, Novosibirsk, 630090 Russia*
[4]*Tomsk State University, Tomsk, 634050 Russia*
[5]*Higher School of Economics, Moscow, 101000 Russia*
*****e-mail: vmefimov@gmail.com*

It is generally accepted that the existing set of proteinogenic amino acids encoded by the standard genetic code was formed step by step in the course of evolution. Most studies name Ala, Asp, Glu, Gly, Ile, Leu, Pro, Ser, Thr and Val as early amino acids, presumably of extraterrestrial origin. However, other studies have chosen a consensus list of early amino acids in which Ile is replaced by Arg. We compared the differences between early and late amino acids for the lists with Ile and with Arg based on their physicochemical properties (AAindex database). The point-biserial correlation coefficient rpb, Student's t-test and its reliability, *p*-value, were calculated between the binary lists with Ile and Arg and each AA index. Since a total of 2×553 *p*-values were obtained, the problem of multiple comparisons was solved using the Bonferroni correction and the Benjamini-Hochberg method. Next, we used the 2B-PLS method, which is applied to two different sets of variables related to the same objects, to find information common to both sets. The first set was the binary lists of Trifonov (Arg) and Wong (Ile), and the second set was 553 AA indexes. The maximum correlation with both the list with Ile and with Arg (1.0 and 0.8, respectively) was demonstrated by the binary AA index CHAM830108, which characterizes the ability of an amino acid to be a charge donor: late amino acids are capable of being donors, while early ones are not. Apparently, this is due to the differences in the conditions under which the standard set of amino acids evolved: prebiotic and biotic. The results of the 2B-PLS analysis also show that in the list of 10 evolutionarily early amino acids, Ile looks preferable to Arg. The allocation of the last 6 amino acids (Cys, His, Met, Phe, Trp, Tyr) obtained on the basis of the reduction of the HOMO-LUMO gap in a separate, third stage of the evolution of the set of standard amino acids is confirmed. A compact arrangement on the 2B-PLS plane of the physicochemical properties of three groups of amino acids, in which adenine, thymine and cytosine are located in the second position of the codons, respectively, as well as the maximum dispersion of amino acids with guanine in the second position of the codons, is revealed.

Keywords: early and late amino acids, AAindex, CHAM830108, point-biserial correlation coefficient, *p*-value, Bonferroni correction, Benjamini-Hochberg method, 2B-PLS analysis